УДК 004.62:004.023

Д. Ю. Тавров, О. Р. Чертов
Національний технічний університет України «Київський політехнічний інститут»
Україна, 03056, м. Київ, пр-т Перемоги, 37

ДВОФАЗОВИЙ МЕМЕТИЧНИЙ АЛГОРИТМ ЗАБЕЗПЕЧЕННЯ ГРУПОВОЇ АНОНІМНОСТІ ДАНИХ

D. Y. Tavrov, O. R. Chertov
National Technical University of Ukraine "Kyiv Polytechnic Institute"
Ukraine, 03056, c. Kyiv, Peremohy ave., 37

TWO-PHASE MEMETIC ALGORITHM FOR PROVIDING DATA GROUP ANONYMITY

Д. Ю. Тавров, О. Р. Чертов
Национальный технический университет Украины «Киевский политехнический институт»
Украина, 03056, г. Киев, пр-т Победы, 37

ДВУХФАЗНЫЙ МЕМЕТИЧЕСКИЙ АЛГОРИТМ ОБЕСПЕЧЕНИЯ ГРУППОВОЙ АНОНИМНОСТИ

У статті розглянуто задачу забезпечення групової анонімності, запропоновано модифікацію меметичного алгоритму її розв'язання, яка передбачає його виконання у дві фази, що дозволяє поліпшити якість розв'язків. Застосування алгоритму продемонстровано за допомогою прикладу на основі реальних даних.
Ключові слова: групова анонімність, меметичний алгоритм, мікрофайл.

In the article, a task of providing group anonymity is discussed. A modification of the memetic algorithm for solving it is presented. The modification implies carrying the algorithm out in two phases, which enables us to obtain higher quality solutions. Application of the algorithm is illustrated with a real data based example.
Key words: group anonymity, memetic algorithm, microfile.

В статье рассматривается задача обеспечения групповой анонимности, предлагается модификация меметического алгоритма ее решения, предусматривающая его исполнение в две фазы, что позволяет улучшить качество решений. Применение алгоритма демонстрируется с помощью примера на основе реальных данных.
Ключевые слова: групповая анонимность, меметический алгоритм, микрофайл.

Вступ

Оскільки людина — істота соціальна, більшість її вчинків залежить від сприйняття тих, чия думка становить для неї найбільшу вагу. Часто особа не готова розкривати інформацію про цих людей, що може пояснюватися суб'єктивними факторами чи природою її оточення (релігійна громада, ЛГБТ-спільнота тощо).

Постає задача приховання належності особи певній групі, що можна сформулювати як задачу маскування певних характеристик особи [1], тобто як задачу забезпечення індивідуальної анонімності, де під анонімністю розуміють властивість суб'єкта бути неідентифіковним у множині інших суб'єктів. Можна поставити пов'язану задачу забезпечення групової анонімності (ЗЗГА), у рамках якої потрібно приховати інформацію не про особу, а про групу осіб (наприклад, замаскувати територіальний розподіл даних про групу військовослужбовців).

Метод забезпечення групової анонімності даних має задовольняти умови [2]:

1. Ризик розкриття інформації після модифікації даних низький.

2. Аналіз початкових та захищених даних повинен давати близькі результати.
3. Вартість модифікації даних прийнятна.

У загальному випадку, забезпечити анонімність можна шляхом вилучення певних атрибутів (наприклад, «Військова служба») із мікрофайлу. Такий підхід найпростіший, проте він задовольняє тільки третю умову з наведених вище. Другу умову він задовольняє меншою мірою: аналіз даних з урахуванням вилученого атрибута стає неможливо провести в принципі. Окрім цього, як показано в [3–4], цей підхід у певних випадках не задовольняє першої умови: існує можливість сформувати так звану нечітку модель групи, за допомогою якої можна оцінити ступінь належності деякого запису групі шляхом аналізу загальних атрибутів («Вік», «Стать» та ін.), вилучення яких неприйнятне. Отже, для повноцінного забезпечення групової анонімності потрібно застосовувати додаткові модифікації даних.

Підходи до розв’язання ЗЗГА можна розділити [5] на ті, що передбачають одержання розв’язку в два етапи, та ті, що передбачають його одержання в один етап. Двоетапні методи передбачають одержання модифікованого розподілу, що маскує чутливі властивості групи, а потім модифікацію первинних даних із метою приведення їх у відповідність до модифікованого розподілу. Якщо на розв’язок ЗЗГА не накладати додаткових обмежень, можна одержати модифіковані розподіли, еквівалентні з погляду маскування чутливих властивостей даних. При цьому різні модифікації вестимуть до одержання на другому етапі спотворень різного обсягу.

Як альтернативу можна використовувати одноетапний підхід до розв’язання ЗЗГА, згідно з яким потрібно одержати модифікований розподіл, який одночасно задовольняє накладені на нього обмеження щодо маскування чутливих властивостей даних і веде до спотворень мінімального обсягу. За такого підходу ЗЗГА є складною задачею умовної оптимізації, і для її розв’язання запропоновано [5] використовувати *меметичні алгоритми* (МА) [6], які, як правило, реалізують у вигляді еволюційних алгоритмів із додаванням процедур локального пошуку [7].

Можна виділити чотири [7–8] способи врахування в еволюційному алгоритмі обмежень, які накладають на розв’язок оптимізаційної задачі:

1. Штрафні функції, що зменшують пристосованість недопустимих розв’язків.
2. Коригуючі функції, що трансформують недопустимі розв’язки в допустимі.
3. Звуження пошуку до підпростору допустимих розв’язків шляхом використання спеціальної схеми кодування для подання особин у популяції.
4. Декодувальні функції, що відображають недопустимі розв’язки на допустимі, трансформуючи таким чином пошуковий простір.

У загальному випадку на розв’язок ЗЗГА не накладають інших обмежень, окрім маскування чутливих властивостей даних, тому доцільним є використання штрафних функцій. Але різні обмеження можуть вести еволюційний процес у різних напрямах, і не завжди в напрямі найменших спотворень.

Метою даної роботи є побудова *двофазового* меметичного алгоритму розв’язання ЗЗГА, на першій фазі якого визначають початкові обмеження та одержують наближені розв’язки, на другій — аналізують ці розв’язки, уточнюють обмеження та одержують остаточні розв’язки.

Задача забезпечення групової анонімності

Нехай дані організовано у вигляді *мікрофайлу* $\mathbf{M}$, записи r_i якого, $i=\overline{1,\rho}$, містять значення атрибутів w_j, $j=\overline{1,\eta}$. Позначимо через $\mathbf{w}_j$ множину значень w_j. Позначмо через w_{v_j}, $j=\overline{1,t}$, *сутнісні атрибути*. *Сутнісну комбінацію значень* можна визначити як елемент із $\mathbf{w}_{v_1}\times\ldots\times\mathbf{w}_{v_t}$. Позначмо через $\mathbf{V}=\{V_1,V_2,\ldots,V_{l_v}\}$ множину цих комбінацій. Записи, значення яких належать $\mathbf{V}$, називатимемо *сутнісними*. Позначмо через w_p *параметризуючий атрибут*, значення якого називатимемо *параметризуючими*. Позначмо через $\mathbf{P}=\{P_1,P_2,\ldots,P_{l_p}\}$ множину параметризуючих значень. За допомогою цих значень можна розбити $\mathbf{M}$ на *підмікрофайли* $\mathbf{M}_1,\ldots,\mathbf{M}_{l_p}$.

Позначмо через $G(\mathbf{V},\mathbf{P})$ *групу* респондентів, дані про яку потрібно захистити. Групу визначають значення параметризуючого та сутнісних атрибутів мікрофайлу.

Задача забезпечення групової анонімності полягає [5] в модифікації $\mathbf{M}$ задля маскування чутливих властивостей даних про групу. Позначмо через $\Omega(\mathbf{M},G)$ *цільове подання* (ЦП), яке подає властивості даних про G у зручний для маскування спосіб. ЗЗГА полягає в підборі перетворення $A:\Omega(\mathbf{M},G)\to\Omega^*(\mathbf{M}^*,G)$, одержанні *модифікованого ЦП* Ω^* і *модифікованого мікрофайлу* $\mathbf{M}^*$. У роботі працюватимемо з ЦП у вигляді *кількісного сигналу* $\mathbf{q}=(q_1,q_2,\ldots,q_{l_p})$, де q_i — число сутнісних записів у $\mathbf{M}_i$, $i=\overline{1,l_p}$. Під чутливими властивостями розумітимемо *викиди* $\mathbf{q}$.

Будь-яке перетворення A повинно забезпечувати два види модифікації даних:

1. Сигнал $\mathbf{q}$ потрібно модифікувати для маскування викидів з урахуванням накладених на його значення обмежень, які на практиці є нечіткими [9].
2. $\mathbf{M}$ потрібно привести у відповідність із модифікованим сигналом шляхом попарного обміну сутнісних та несутнісних записів між підмікрофайлами, і при цьому повинно бути внесено спотворення мінімального обсягу.

Обсяг спотворень оцінюють за допомогою *визначальної метрики* [5]:

$$\mathrm{InfM}\left(r,r^*\right)=\sum_{p=1}^{n_{\text{пор}}}\omega_p\left(\frac{r\left(I_p\right)-r^*\left(I_p\right)}{r\left(I_p\right)+r^*\left(I_p\right)}\right)^2+\sum_{k=1}^{n_{\text{кат}}}\gamma_k\chi^2\left(r\left(J_k\right),r^*\left(J_k\right)\right), \quad (1)$$

де I_p (J_k) — p-ий порядковий (k-ий категорійний) *визначальний атрибут* (атрибут, розподіл значень якого становить інтерес для дослідників), $r(\cdot)$ повертає значення вказаного атрибута запису r, $\chi(v_1,v_2)$ — оператор, що дорівнює χ_1, якщо v_1 та v_2 належать одній категорії, та χ_2 — інакше, ω_p та γ_k — невід'ємні ваги, які підбирають, виходячи з важливості атрибута (що він важливіший, то більша вага).

Меметичний алгоритм

Розглянемо перетворення A у вигляді наступного меметичного алгоритму [5]:

1. Випадковим чином згенерувати популяцію $P=\{U_i\}$ з μ особин, $i=\overline{1,\mu}$.

2. Застосувати оператор локального пошуку $S(U_i)\ \forall i=\overline{1,\mu}$.

3. Обчислити значення функції пристосованості $f(U_i)\ \forall i=\overline{1,\mu}$.

4. Якщо виконується умова завершення, зупинити алгоритм.

5. Вибрати λ пар батьківських особин; помістити їх у множину P'.

6. Застосувати оператор рекомбінації $R\left(U_{i_1},U_{i_2}\right)$ до кожної пари особин $\left\langle U_{i_1},U_{i_2}\right\rangle$ з P', $i_1=\overline{1,\lambda}$, $i_2=\overline{1,\lambda}$, $i_1\neq i_2$; помістити нащадків у множину P''.

7. Застосувати оператор мутації $M\left(U_j\right)\ \forall U_j\in P''$, $j=\overline{1,\lambda}$.

8. Застосувати $S\left(U_j\right)$ до кожної особини з P'', $j=\overline{1,\lambda}$.

9. Обчислити значення функції пристосованості $f\left(U_j\right)\ \forall j=\overline{1,\lambda}$.

10. Вибрати μ пристосованіших особин із $P\cup P''$; додати у P замість поточних.

11. Перейти на крок 3.

Кожна особина в P є матрицею U з Q рядками та 4 стовпцями:

1. Елементи першого (третього) стовпця $u_{i1}\ \ \forall i=\overline{1,Q}$ відповідають індексам підмікрофайлів, із яких потрібно вилучити (у які потрібно додати) сутнісні записи.

2. Елементи другого стовпця u_{i2} визначають індекси записів із $\mathbf{M}_{u_{i1}}$, які потрібно вилучити. Елементи четвертого стовпця u_{i4} визначають індекси записів із $\mathbf{M}_{u_{i3}}$, які потрібно обміняти з записами, визначеними u_{i2}.

Кожна особина U однозначно визначає розв’язок ЗЗГА.

У даній роботі пропонується використовувати функцію пристосованості

$$f(U)=\Upsilon(U)\cdot\Phi(U)\cdot\Psi(U),\tag{2}$$

де $\Upsilon(U)$ — оцінка розв’язку з погляду мінімізації обсягу спотворень, $\Phi(U)$ — штрафна функція — оцінка розв’язку з погляду маскування викидів, $\Psi(U)$ — штраф, уведений для упередження необмеженого збільшення числа рядків в особинах. Значення кожного множника з (2) повинні лежати в проміжку $[0,1]$.

Як $M(U)$ доцільно використовувати оператор, який є суперпозицією $M=M_4\circ M_3\circ M_2\circ M_1$ операторів, що діють на кожний стовпець U окремо.

Двофазовий меметичний алгоритм

Для деякого елемента $\mathbf{q}$ можна визначити обмеження одного з двох типів:

1. *Спадне обмеження*, функція належності якого монотонна незростаюча, що прямує до 1 зі спадом значення елемента до заданого порогового значення.

2. *Зростаюче обмеження*, функція належності якого монотонна неспадна, що прямує до 1 зі зростанням значення елемента до заданого порогового значення.

На початку розв’язання ЗЗГА можна визначити тільки спадні обмеження для підмікрофайлів, із яких потрібно вилучити записи. Вибір підмікрофайлів для додавання записів (та відповідних зростаючих обмежень) є неоднозначним, і його можна перенести на еволюційний процес відповідно до такої процедури:

1. На підставі аналізу $\mathbf{q}$ сформувати спадні обмеження для тих його елементів, які порушують вимогу щодо маскування викидів сигналу.
2. Виконати МА.
3. Розділити одержані особини на *допустимих* (сумісні з обмеженнями та маскують викиди), *майже допустимі* (сумісні з обмеженнями, але не маскують викидів) та *недопустимі* (не сумісні з обмеженнями).
4. Згрупувати в кластери всі майже допустимі особини, для яких можна задати однакові зростаючі обмеження (одна особина може належати декільком кластерам).
5. Вибрати кластер із найменшим середнім значенням (1); якщо він містить менше μ особин, збільшити його розмір до μ випадковим копіюванням особин; якщо більше μ особин, зменшити його розмір до μ випадковим видаленням особин.
6. Застосувати МА до множини особин, одержаної на кроці 5.

Перші два кроки становлять *першу фазу* МА, решта чотири — *другу фазу* МА.

Методика забезпечення групової анонімності

Розглянемо методику забезпечення групової анонімності, яка, окрім забезпечення анонімності у спосіб, описаний вище, враховує випадок забезпечення анонімності груп, відносно яких існує загроза її порушення шляхом аналізу значень несутнісних атрибутів мікрофайлу. В останньому випадку анонімність групи можна порушити за допомогою її нечітких моделей: на основі сторонніх даних [3] або на основі експертних знань [4]. Методику можна застосовувати, якщо дані мікрофайлу зберігаються в текстовому файлі, де кожний рядок відповідає певному респонденту, а значення в рядку відповідають значенням атрибутів цього респондента.

Методику забезпечення групової анонімності можна розбити на етап побудови моделі групи, етап побудови нечіткої моделі групи на основі сторонніх даних, етап побудови нечіткої моделі групи на основі експертних знань, етап розв'язання ЗЗГА.

На *етапі побудови моделі групи* потрібно вибрати мікрофайл $\mathbf{M}$, визначити сутнісні та параметризуючі атрибути, обчислити значення ЦП та визначити викиди.

На *етапі побудови нечіткої моделі групи на основі сторонніх даних* потрібно за допомогою моделі перевірити, чи має місце загроза порушення анонімності у випадку вилучення з $\mathbf{M}$ сутнісних атрибутів. Для цього потрібно вибрати допоміжний мікрофайл $\tilde{\mathbf{M}}$ [3] та виконати гармонізацію $\mathbf{M}$ та $\tilde{\mathbf{M}}$ з одержанням гармонізованих $\mathbf{M}^H$ та $\tilde{\mathbf{M}}^H$ з ідентичною структурою атрибутів. Якщо $\tilde{\mathbf{M}}$ вибрати неможливо, потрібно перейти на наступний етап методики. Наступною є ідентифікація вхідних змінних системи нечіткого виведення [3], що є основою нечіткої моделі, а також побудова її бази правил. Після цього потрібно збудувати ЦП, що відповідає нечіткій моделі, та визначити в ньому викиди. Якщо викиди відповідатимуть викидам початкового ЦП, існує загроза порушення анонімності. Якщо збудувати базу правил неможливо, можна перейти на етап розв'язання ЗЗГА.

На *етапі побудови нечіткої моделі групи на основі експертних знань* потрібно за допомогою моделі перевірити, чи має місце загроза порушення анонімності у випадку вилучення з $\mathbf{M}$ сутнісних атрибутів. Після побудови моделі згідно з [4] потрібно побудувати ЦП, що їй відповідає, та визначити його викиди. Якщо викиди відповідатимуть викидам початкового ЦП, існує загроза порушення анонімності.

На *етапі розв'язання ЗЗГА* в різних випадках потрібно використовувати різні ЦП: у випадку загрози порушення анонімності за допомогою однієї з нечітких моделей групи — ЦП, що відповідає цій моделі; у випадку відсутності такої

загрози — початкове ЦП. Якщо з $\mathbf{M}$ вирішено не вилучати сутнісних атрибутів, ЗЗГА додатково потрібно розв’язати для початкового ЦП незалежно від того, чи існує загроза порушення анонімності за допомогою будь-якої з нечітких моделей.

Для оцінювання якості розв’язків ЗЗГА з погляду маскування викидів ЦП потрібно сформувати спадні обмеження. Для оцінювання якості розв’язків ЗЗГА з погляду мінімізації внесених спотворень потрібно визначити параметри метрики (1).

Для розв’язання ЗЗГА потрібно застосувати описаний вище МА. По завершенню його роботи потрібно відібрати з числа особин останнього покоління допустимі. Якщо таких не виявлено або їхня якість незадовільна, потрібно виконати другу фазу МА. Якщо після її виконання розв’язок одержати неможливо, МА варто перезапустити, змінивши окремі його параметри. Після відбору розв’язку потрібно виконати модифікацію $\mathbf{M}$, яка передбачає виконання фізичного обміну записів $\mathbf{M}$, а також запис даних модифікованого мікрофайлу у файл відповідного формату.

Практичні результати

Розгляньмо задачу маскування територіального розподілу військовослужбовців штату Массачусетс, США. Як початкові дані, було взято мікрофайл перепису населення США 2000 р. [10], що містить 141 838 записів. Вважатимемо, що загроза порушення анонімності за допомогою нечіткої моделі відсутня.

Для визначення групи як сутнісний атрибут було взято «Військову службу», сутнісне значення — «1» (відповідає «Поточній службі»), параметризуючий атрибут — «Місце роботи», значеннями якого є коди статистичних областей штату Массачусетс, параметризуючі значення — кожне десяте значення від 25010 до 25120 (ці значення відповідають кодам статистичних областей штату Массачусетс).

Кількісний сигнал $\mathbf{q}$ представлено на рисунку 1. Елементи 1–12 відповідають областям 25010–25120. Анонімність можна забезпечити шляхом зменшення 2, 7, 9 та 12 значень сигналу, тобто на відповідні значення модифікованого сигналу $\mathbf{q}^*$ потрібно накласти спадні обмеження, які характеризуються наступними функціями належності: $\mu_2(x)=ZMF(x,20,67)$, $\mu_7(x)=ZMF(x,25,30)$, $\mu_9(x)=ZMF(x,25,28)$,

$$\mu_{12}(x)=ZMF(x,25,38), \text{ де } ZMF(x,a,b)=\begin{cases}1, & x\le a\\ 1-2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2}\\ 2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b\\ 0, & x\ge b\end{cases}.$$

Порогові значення обмежень обрано з метою зменшення значень відповідних елементів сигналу до рівня, зіставного з величиною найменшого викиду, яким є 7-ий елемент сигналу. Індекси 2, 7, 9 та 12 було вибрано як індекси, що можуть входити в 1-ий стовпець особин у популяції в МА, інші індекси — у 3-ій стовпець.

Для мінімізації обсягу внесених спотворень як визначальні атрибути було взято «Стать», «Вік», «Іспанське чи латиноамериканське походження», «Сімейний стан», «Рівень освіти», «Громадянство», «Сукупний дохід». Кожний атрибут вважався категорійним. Для спрощення інтерпретації (1) було вибрано наступні її параметри:

$\gamma_k = 1 \ \forall k = \overline{1,7}$, $\chi_1 = 1$, $\chi_2 = 0$. Метрика (1) у цьому разі показує кількість значень атрибутів, які потрібно модифікувати за один обмін записів між підмікрофайлами.

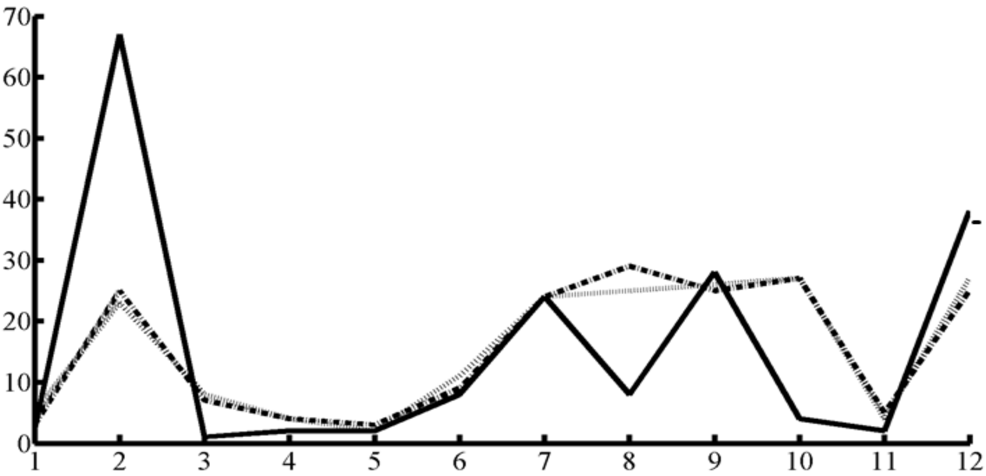


Рис. 1. Початковий кількісний сигнал (суцільна лінія), модифікований сигнал з метрикою 37 (штрих-пунктирна), модифікований сигнал з метрикою 38 (пунктирна)

Функція пристосованості (2) першої фази алгоритму має наступний вигляд:

$$f^{ex_1}(U) = \frac{1099 - \sum_{i=1}^{Q}\sum_{k=1}^{7} \operatorname{sign}\left|\mathbf{M}_{u_{i1}}(u_{i2}, W_k) - \mathbf{M}_{u_{i3}}(u_{i4}, W_k)\right|}{1099} + \prod_{j \in J} \mu_j\left(q_j^*(U)\right) + \frac{1}{1 + e^{0.5(Q_U - 90)}},$$

де W_k — k-ий визначальний атрибут, $k = \overline{1,7}$, $\mathbf{M}_j(i, W_k)$ — оператор, який повертає значення атрибута W_k i-го запису підмікрофайлу $\mathbf{M}_j$, останній доданок — штрафний терм, який дискримінує особин із кількістю рядків, більшою за 100, $C_{\max}$ — найбільше можливе сумарне значення (1), обчисленої для всіх пар в U.

Як оператор рекомбінації було вибрано оператор, описаний у [5]. Як оператори мутації M_1 та M_2 було обрано мутацію обміну [11], M_3 та M_4 — мутацію випадкової заміни [7, с. 43]. Як локальний пошук було вибрано оператор, описаний у [12]. Як метод відбору було вибрано турнірний відбір [13], із турніром розміру 5.

Першу популяцію було ініціалізовано шляхом випадкової генерації матриць із різною кількістю рядків. Елементи 1-го стовпця генерувалися з імовірностями, пропорційними відповідним значенням $\mathbf{q}$, 3-го — імовірностями, пропорційними числу записів у відповідних підмікрофайлах. Інші параметри МА було обрано так: $\mu = 100$, $\lambda = 40$, імовірність рекомбінації $p_c = 1$, $p_{m_1} = p_{m_2} = p_{m_3} = p_{m_4} = 0{,}001$, параметр локального пошуку (див. [12]) $p_{mem} = 0{,}75$. Для упередження передчасної збіжності МА імовірність мутації збільшувалася вдесятеро щоразу, коли середньоквадратичне відхилення пристосованостей ставало меншим за 0,03.

Було виконано 30 запусків МА. МА припиняв роботу після генерації 1000 популяцій. Серед 3000 особин, одержаних за результатами першої фази, 754 особини (25,133%) є допустимими. Середнє значення сумарної метрики (1) по всіх

допустимих особинах дорівнює 57,901. Більшість особин майже допустимі (1837, або 61,233%). Їх було розбито на кластери, найбільші з яких наведено в таблиці 1.

Таблиця 1. Кластери, одержані після першої фази МА

Елементи сигналу для збільшення	Розмір кластера	Середня метрика
1 та 6	78	45,436
3 та 6	84	46,048
3 та 10	26	46,269
4 та 6	43	48,488
6 та 8	183	46,519
8 та 10	101	44,238

Для другої фази доцільно вибрати особини з останнього кластеру. Можна сформувати обмеження, що зростають, з наступними функціями належності:

$$\mu_8(x)=\mu_{10}(x)=\begin{cases}1, & x\le 15\\ 2\left(\dfrac{x-15}{12}\right)^2, & 15\le x\le 21\\ 1-2\left(\dfrac{x-27}{12}\right)^2, & 21\le x\le 27\\ 0, & x\ge 27\end{cases}.$$

Порогові значення функцій вибрано так, щоб у модифікованих сигналах, сумісними з обмеженнями з високим ступенем, початкові викиди було масковано (8-е та 10-е значення стануть співмірними зі значеннями з індексами 2, 7, 9 та 12).

Функція пристосовності (2) другої фази алгоритму має наступний вигляд:

$$f^{ex_2}(U)=\frac{1099-\sum_{i=1}^{Q}\sum_{k=1}^{7}\operatorname{sign}\left|\mathbf{M}_{u_{i1}}(u_{i2},W_k)-\mathbf{M}_{u_{i3}}(u_{i4},W_k)\right|}{1099}+\prod_{j\in J\cup\{10,12\}}\mu_j\left(\left|U\right|^{(j)}\right)+\frac{1}{1+e^{0.5(Q_U-90)}},$$

Серед 3000 особин 2693 (89,767%) є допустимими. Два розв’язки з найменшою сумарною метрикою (1) представлено на рисунку 1. Середнє значення метрики по допустимих особинах дорівнює 47,873, тобто для анонімізації достатньо змінити не більше за 0,005% значень атрибутів мікрофайлу. Ці результати ліпші від одержаних у [5], що свідчить про ефективність двофазового підходу до розв’язання ЗЗГА.

Висновки

У роботі запропоновано модифікацію меметичного алгоритму розв’язання задачі забезпечення групової анонімності, яка передбачає виконання алгоритму у дві фази: на першій фазі формуються початкові обмеження на розв’язок задачі, а на другій їх уточнюють з урахуванням результатів виконання першої фази.

Застосування двофазового алгоритму до реальних даних свідчить, що для анонімізації даних мікрофайлу достатньо змінити не більше за 0,005% значень його атрибутів. Цей результат ліпший від одержуваного за допомогою однофазового алгоритму, що підтверджує практичну корисність запропонованої модифікації.

Література

1. Fung B. Privacy-preserving data publishing: a survey of recent developments / B. Fung, K. Wang, R. Chen, P. Yu // ACM Computing Surveys. — 2010. — 42(4). — P. 1–53.
2. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // International Symposium on Computing, Communication and Control (ISCCC 2009). Proc. of CSIT, vol. 1. — Singapore : IACSIT Press, 2011. — P. 339–343.
3. Чертов О. Р. Эволюционный алгоритм построения нечеткой модели группы с целью нарушения ее анонимности / О. Р. Чертов, Д. Ю. Тавров // Международная научная конференция имени Т. А. Таран «Интеллектуальный анализ информации» ИАИ-2015, Киев, 20–22 мая 2015 г. : сб. тр. / гл. ред. С. В. Сирота. — К. : Просвіта, 2015. — С. 272–280/
4. Chertov O. Microfiles as a Potential Source of Confidential Information Leakage / O. Chertov, D. Tavrov // Intelligent Methods for Cyber Warfare [ed. R. R. Yager, M. Z. Reformat, N. Alajlan]. — Springer International Publishing Switzerland, 2015. — P. 87–114.
5. Chertov O. Memetic Algorithm for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Advance Trends in Soft Computing [ed. M. Jamshidi, V. Kreinovich, J. Kacprzyk]. — Springer International Publishing Switzerland, 2014. — P. 281–292.
6. Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: Toward memetic algorithms / Pablo Moscato // C3P Report 826. — Caltech, CA, 1989. — P. 33–48.
7. Eiben A. E. Introduction to Evolutionary Computing / A. E. Eiben, J. E. Smith. — Berlin, Heidelberg : Springer-Verlag, 2007. — 316 p.
8. Evolutionary Computation 2. Advanced Algorithms and Operators [ed. T. Bäck, D. B. Fogel, Z. Michalewicz]. — Bristol, Philadelphia : Institute of Physics Publishing, 2000. — 308 p.
9. Чертов О. Р. Меметичний алгоритм із нечіткими обмеженнями для розв’язання задачі забезпечення групової анонімності / О. Р. Чертов, Д. Ю. Тавров // Інформаційна безпека. — 2013. — №4 (12). — С. 135–144.
10. Census 2000. 5-Percent Public Use Microdata Sample Files [Електронний ресурс]. — Режим доступу: http://www.census.gov/main/www/cen2000.html.
11. Syswerda G. Schedule optimization using genetic algorithms / G. Syswerda // Handbook of Genetic Algorithms [ed. L. Davis]. — New York : Van Nostrand Reinhold, 1991. — P. 332–349.
12. Чертов О. Р. Меметичний алгоритм для модифікації мікрофайлу з мінімізацією спотворень у процесі забезпечення групової анонімності / О. Р. Чертов, Д. Ю. Тавров // Штучний інтелект. — 2013. — №3 (61). — С. 399–410.
13. Brindle A. Genetic algorithms for function optimization : [doctoral dissertation and tech. rep. TR81-2] / A. Brindle. Edmonton : University of Alberta, Department of Computer Science, 1981. — 93 p.

Literatura

1. Fung B. Privacy-preserving data publishing: a survey of recent developments / B. Fung, K. Wang, R. Chen, P. Yu // ACM Computing Surveys. — 2010. — 42(4). — P. 1–53.
2. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // International Symposium on Computing, Communication and Control (ISCCC 2009). Proc. of CSIT, vol. 1. — Singapore : IACSIT Press, 2011. — P. 339–343.
3. Chertov O. R. Evolutionary algorithm for constructing fuzzy model of a group in order to violate its anonymity / O. R. Chertov, D. Y. Tavrov // T. A. Taran International scientific conference “Intelligent Analysis of Information” IAI-2015, Kiev, May 20–22, 2015 : proceedings / ed. S. V. Syrota. — K. : Prosvita, 2015. — S. 272–280.
4. Chertov O. Microfiles as a Potential Source of Confidential Information Leakage / O. Chertov, D. Tavrov // Intelligent Methods for Cyber Warfare [ed. R. R. Yager, M. Z. Reformat, N. Alajlan]. — Springer International Publishing Switzerland, 2015. — P. 87–114.
5. Chertov O. Memetic Algorithm for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Advance Trends in Soft Computing [ed. M. Jamshidi, V. Kreinovich, J. Kacprzyk]. — Springer International Publishing Switzerland, 2014. — P. 281–292.
6. Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: Toward memetic algorithms / Pablo Moscato // C3P Report 826. — Caltech, CA, 1989. — P. 33–48.
7. Eiben A. E. Introduction to Evolutionary Computing / A. E. Eiben, J. E. Smith. — Berlin, Heidelberg : Springer-Verlag, 2007. — 316 p.
8. Evolutionary Computation 2. Advanced Algorithms and Operators [ed. T. Bäck, D. B. Fogel, Z. Michalewicz]. — Bristol, Philadelphia : Institute of Physics Publishing, 2000. — 308 p.
9. Chertov O. R. Memetic algorithm with fuzzy restrictions for solving the task of providing group anonymity / O. R. Chertov, D. Y. Tavrov // Informatsiina Bezpeka. — 2013. — №4 (12). — S. 135–144.

10. Census 2000. 5-Percent Public Use Microdata Sample Files [Elecronic resource]. — Mode of access: http://www.census.gov/main/www/cen2000.html.
11. Syswerda G. Schedule optimization using genetic algorithms / G. Syswerda // Handbook of Genetic Algorithms [ed. L. Davis]. — New York : Van Nostrand Reinhold, 1991. — P. 332–349.
12. Chertov O. R. Memetic algorithm for microfile modification with minimizing distortion while providing group anonymity / O. R. Chertov, D. Y. Tavrov // Shtuchnyi Intellekt. — 2013. — №3 (61). — S. 399–410.
13. Brindle A. Genetic algorithms for function optimization : [doctoral dissertation and tech. rep. TR81-2] / A. Brindle. Edmonton : University of Alberta, Department of Computer Science, 1981. — 93 p.

RESUME

D. Y. Tavrov, O. R. Chertov

Two-Phase Memetic Algorithm for Providing Data Group Anonymity

Nowadays, it has become a common practice to provide public access to various kinds of primary non-aggregated statistical data. Necessary precautions ought to be taken to guarantee that sensitive data features are masked, and data privacy cannot be violated.

In case of protecting information about a group of people, it is important to protect intrinsic data features. To do so, it is obligatory to introduce a certain level of distortion into the data. Minimizing this distortion is a complex optimization task, which can be solved by applying appropriate heuristic techniques, e.g., memetic algorithms.

In the paper, we propose a modification of the memetic algorithm for solving the task of providing group anonymity. The modified algorithm consists of two phases, where on the first phase initial constraints on the solution are stated, and on the second phase they are refined using the data obtained as the result of the first phase.

We illustrate the application of the two-phase algorithm by solving a task of providing group anonymity based on real data.

Д. Ю. Тавров, О. Р. Чертов

Двухфазный меметический алгоритм обеспечения групповой анонимности

На сегодняшний день все чаще предоставляют публичный доступ к различным первичным неагрегированным статистический данным. При этом необходимо предпринимать меры для маскирования чувствительных к раскрытию особенностей данных с тем, чтобы предотвратить нарушение приватности данных.

В случае защиты информации о группе лиц, важно защищать присущие им особенности. Этого невозможно достичь без внесения в данные искажений определенного объема. Минимизация таких искажений является сложной оптимизационной задачей, для решения которой можно применять соответствующие эвристические методы, например, меметические алгоритмы.

В статье предлагается модификация меметического алгоритма решения задачи обеспечения групповой анонимности. Модифицированный алгоритм состоит из двух фаз, при этом на первой фазе формируются начальные ограничения на решения, а на второй они уточняются на основе данных, полученных по результатам первой фазы.

Применение двухфазного алгоритма проиллюстрировано путем решения задачи обеспечения групповой анонимности, основанной на реальных данных.

Надійшла до редакції 26.08.2015